

Martin Nilsson

*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501 USA**Institute of Theoretical Physics, Chalmers University of Technology and Göteborg University, S-412 96 Göteborg, Sweden*`martin@fy.chalmers.se`

Nigel Snoad

*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501 USA**The Australian National University, ACT 0200, Australia `nigel@santafe.edu`*

(April 21, 2000)

In this paper we study the evolution of the mutation rate for simple organisms in dynamic environments. A model with multiple fitness coding loci tracking a moving fitness peak is developed and an analytical expression for the optimal mutation rate is derived. Surprisingly it turns out that the optimal mutation rate per genome is approximately independent of genome length, something that also has been observed in nature. Simulations confirm the theoretical predictions. We also suggest an explanation for the difference in mutation frequency between RNA and DNA based organisms.

I. INTRODUCTION

In any given environment the vast majority of mutations that have any effect on the fitness of a biological organism are deleterious. One might expect the damaging effect of non-zero mutation rates to imply that when under evolutionary control the lowest mutation rate compatible with physiological constraints should be selected for. However, when examined experimentally bacteria and viruses (and indeed all organisms) have significant non-zero rate, the magnitude and diversity of which have failed to find satisfactory theoretical explanation. Some results from a number of experiments measuring the mutation rates of a selection of small DNA-based organisms are shown in Table I.

Organism	ν	μ_b	μ_G
Bacteriophage <i>M13</i>	$6.4 \cdot 10^3$	$7.2 \cdot 10^{-7}$	0.0046
Bacteriophage <i>l</i>	$4.9 \cdot 10^4$	$7.7 \cdot 10^{-8}$	0.0038
Bacteriophage <i>T2</i> & <i>T4</i>	$1.7 \cdot 10^5$	$2.4 \cdot 10^{-8}$	0.0040
<i>E. coli</i>	$4.6 \cdot 10^6$	$4.1 \cdot 10^{-10}$	0.0025
<i>S. cerevisiae</i>	$1.2 \cdot 10^7$	$2.2 \cdot 10^{-10}$	0.0027
<i>N. crassa</i>	$4.2 \cdot 10^7$	$7.2 \cdot 10^{-11}$	0.0030

TABLE I. Spontaneous mutation rates (per base μ_b and per genome μ_G) in DNA-based microbes with different genome lengths ν . (Data reproduced from Drake *et al.* [1])

Despite the huge variation in genome length over four orders of magnitude the mutation rate per genome and replication μ_G remains constant roughly within a factor of roughly 2 (which is at the same level as the estimated accuracy of the figures). As pointed out by Drake and others [1,2] this constancy in μ_G is surprising since DNA/RNA repair and transcription are primarily local processes that act on individual bases. Thus the data strongly suggest that point mutation rates for the different organisms have evolved towards individual optimal values that result in almost constant genomic copying fidelity.

In this paper we develop a model of the evolution of mutation rates based on changing environments. The evolved point mutation rate of this model scales so that the genomic copying fidelity is approximately independent of genome length and insensitive to other parameters in the model. The evolved mutation rates are also of the same magnitude as observed in Table I for biologically plausible parameter settings. We also suggest a possible explanation for the high mutation rates of RNA viruses. Simulations confirm the predictions of the model.

II. EVOLVING MUTATION RATES

It is impossible to perfectly maintain and copy genetic information. All molecules, including DNA and RNA are thermodynamically unstable, and their physical structure and hence the information they encode changes over time. In addition the binding sites of enzymes such as DNA polymerase are not perfectly specific and errors will be introduced during replication. Lowering the error rate requires the use of increasingly complex proof-reading and repair mechanisms, all of which ultimately impose an energetic, and hence fitness, cost on the organism. We can expect a balance to develop between the pressure to lower mutation rates due to the fitness cost of deleterious mutants and the physiological cost of high copying accuracy [3–5]. Such a balance certainly provides an ultimate lower limit to the mutation rate of all organisms but explaining the constancy in genomic copying fidelity using such arguments causes unnatural assumptions on the relation between cost of local copying fidelity and genome

length. There is also little experimental evidence that mutation rates are actually determined by such a balance.

When viewed as a whole the genome encodes not only proteins that directly influence its reproductive or survival ability, but also the copying fidelity with which the genome reproduces. For example some viroids contain genes that are translated into surface coat proteins while others genes code for the replicase enzymes that perform the copying of its genetic material. In more complex organisms additional genes may encode for modifiers of the accuracy of copy and repair enzymes, usually increasing mutation rates [6–9], but sometimes resulting in a decrease [10]. These modifiers can have large or small effects on mutation rate and affect individual bases or the entire genome [11–13].

One consequence of this flexibility of mutation rates and their encoding is that if there are random changes (mutations) in genes determining the mutation rate then the copying fidelity will itself undergo Darwinian evolution.

III. POPULATION GENETICS IN CHANGING ENVIRONMENTS

When comparing two haploid genomes, the one with lower mutation frequency will produce offspring that are on average more closely related to itself. This means that for an asexual haploid replicator evolving on a static fitness landscape the optimal mutation rate for a sequence whose fitness is already globally maximal is zero. If the fitness peak moves, however, the situation changes: to avoid extinction a genome with an initially superior fitness is forced to accept a non-zero mutation rate to survive. This leads to a non-trivial optimal copying fidelity.

Kimura formalized the evolutionary effect of a changing environment by considering the genetic load of a population [3]: the proportion by which the population fitness is decreased in comparison with an optimum genotype. Genetic load results from a number of competing factors; most notably the *mutational load* due to the deleterious effects of most mutations and the *segregational load* due to the temporary reduction in fitness that occurs whenever the selective environment changes. Assuming that a population minimizes the genetic load, the optimal mutation rate can be calculated. Using a discrete time model, i.e. a model where there is no overlap between generations, with one fitness determining locus the optimal mutation rate becomes:

$$\mu_{opt} = \frac{1}{\tau} \quad (1)$$

where τ is the number of generations between environmental changes. This model only considers the effect of mutations on the population and is therefore based on group selection.

Later population genetic models that examined competition between genetic modifiers of the mutation rate demonstrated that (for haploids with a single fitness determining locus) a non-zero mutation rate comes to dominate a population evolving in an oscillating environment [14–17]. These models are not built on group selection. However a general and simple to interpret multi-locus modifier model does not exist.

IV. THE MODEL

We will explore a more general model of the evolution of mutation rates in a dynamic environment. Consider a population of haploid genomes where a genome consists of two separated parts, one coding for the fitness and one coding for the probability per base μ of an error occurring during copying. There is complete linkage (no recombination) between the sections of the genome that encode the mutation rate and those that determine the fitness. We also assume that the fitness determining region is of fixed length ν . In general we are interested in the fates of certain genomes g_i which have a (possibly time-dependent) fitness advantage $\sigma(t)$ over all other sequences. We call these genomes master-sequences. The genomic copying fidelity of the fitness determining region of each strain g_i is $Q_i = (1 - \mu_i)^\nu$, the index i refers to the mutation rate of the strain, different strains have different mutation rates but identical fitness σ . We assume that mutations do not affect the copying fidelity, only the fitness. Changes to the mutation rates occur on a time-scale significantly slower than the time it takes for the population to reach equilibrium. During a period when a specific sequence has superior fitness compared to the background (i.e. between environmental shifts) the changes in the relative concentrations x_i of the master-sequences are described by the replicator equation

$$\dot{x}_i(t) = Q_i \sigma(t) x_i(t) - f(t) x_i(t) \quad (2)$$

where $f(t) = \sigma(t) \sum_j Q_j x_j(t)$ normalizes the relative concentrations of the master-sequence strains. Mutations from background sequences onto the strains with optimal fitness are ignored. Since we are only interested in competition between master-sequences the background is not explicitly expressed in these equations.

The environment changes as follows: for a time $t \in [0, \tau_1]$ one genotype has superior fitness, followed by a new gene-sequence for time $t \in [\tau_1, \tau_1 + \tau_2]$, etc. The notation is chosen so that τ denotes lengths of time intervals. We assume that the initial concentration of the new master-sequences x_i immediately after the shift (at time $t_a = \sum_{i=1}^m \tau_i + \epsilon$, where m denotes shifts of the fitness-peak and ϵ is an infinitely small time-period) are proportional to the concentrations of the old master-sequence before the shift (at $t_b = \sum_{i=1}^m \tau_i - \epsilon$)

$$x_i(t_a) = h(\mu_i) x_i(t_b) \quad (3)$$

It is reasonable to assume that $h(\mu_i)$ is a function with Taylor-expansion in the mutation rate μ

$$h(\mu) = \sum_{j=k_m}^{\infty} \alpha_j \mu^j \quad (4)$$

where k_m is a measure of the environmental change, i.e. the number of point mutations needed to transform the old superior sequence into the new. This basically means that k_m is the Hamming distance from the old peak to the new at shift m . The constants α_j are combinatorial factors. It will turn out that the optimal mutation rate is independent of these factors.

To analyze the long term behavior of this system we make a change of variables $y_i(t) = e^{\int_0^t f(s)ds} x_i(t)$. The new system of differential equations is linear and the equations are decoupled (due to the assumption that the selective dynamics is significantly faster than the changes in mutation rate), it is therefore easy to find the analytical solution:

$$y_i(t) = y_i(0) e^{Q_i \int_0^t \sigma_m(s) ds} \quad (5)$$

Since x_i is proportional to y_i , maximizing the growth of y_i and x_i are equivalent. After a suitably long time interval the population will be completely dominated by genomes that have a mutation rate closest to the optimal value μ_{opt} which maximizes the long term growth of the strain

$$\max_{\mu} \left(\Pi_m h(\mu) e^{(1-\mu)^{\nu} \langle \sigma \rangle_m \tau_m} \right) \quad (6)$$

where $\langle \cdot \rangle_m$ denotes a time average during time-period m . Setting the derivative of this expression to zero and using Eq. 4 we find the optimal copying fidelity to be approximately

$$\mu_{opt} = \frac{\langle k \rangle}{\nu \langle \sigma \rangle \langle \tau \rangle} \quad (7)$$

where $\langle \cdot \rangle$ denotes a time average over all time periods. We also assume no correlation between $\langle \sigma \rangle_m$ and τ_m . Since the genome lengths is large $\nu \gg 1$, the optimal copying fidelity and mutation rate per genome become:

$$Q_{opt} = e^{-\frac{\langle k \rangle}{\langle \sigma \rangle \langle \tau \rangle}} \quad (8)$$

$$\mu_G = \frac{\langle k \rangle}{\langle \sigma \rangle \langle \tau \rangle} \quad (9)$$

Thus we find that the genomic optimal copying fidelity is independent of the genome length for fairly general types of environmental change in both the advantage of the fittest genotype $\sigma(t)$ and the size of environmental shifts $h(\mu)$.

V. SIMULATIONS

To confirm the theoretical derivations we simulated the evolution of replicators in continuous time on a moving single peaked landscape using a birth-death process. Each time unit in the continuous time replicator equation is the mean replacement time of the population and could therefore be identified as a generation. In the simulation each generation is divided into N time-steps (where N is the population size). At each of these time-steps a single individual is selected to copy and mutate. Individuals are selected with a probability proportional to their relative fitness, which is given by σ or 1 on the single-peaked landscape. Thus a master-sequence of strain g_i (with mutation rate μ_i) is chosen with probability $\frac{x_i \sigma}{\langle f \rangle}$. This copy replaces a randomly chosen individual in the existing population which is then discarded. Thus the population is replaced one by one in discrete birth-death events. In the limit of large population size the dynamics of this simulation approaches the continuous time replicator equation.

The fitness peak is changed every τ generations to one of its nearest neighbors. For the binary genomes used here it is accomplished by flipping a randomly chosen bit in the definition of the fitness peak.

The population was first seeded with a diverse range of mutation rates and the population was allowed to evolve while these rates were kept fixed. This is a true test of μ_{opt} , since the fastest growing sequence should come to dominate. In general the population converged to the strain with mutation rate closest to the theoretically predicted μ_{opt} . Figure 1 shows the mean mutation rate of the population $\bar{\mu}$ evolving down towards the theoretically predicted optimum $\mu_{opt} \approx \frac{1}{\nu \sigma \tau} = 0.00445$. From about generation 800 the variance in mutation rates in the population is larger than the fluctuations in the mean and the evolution of rates has effectively ended.

Simulations were also made to study the effects of more rapidly changing mutator dynamics. In these simulations errors in the copying process not only introduce changes in the fitness determining genotype, but also result in offspring with slightly different mutation rates than their parents, i.e. the mutation rate is allowed to evolve. The mutation rate was treated as a continuous variable which had Gaussian noise introduced during the copying process.

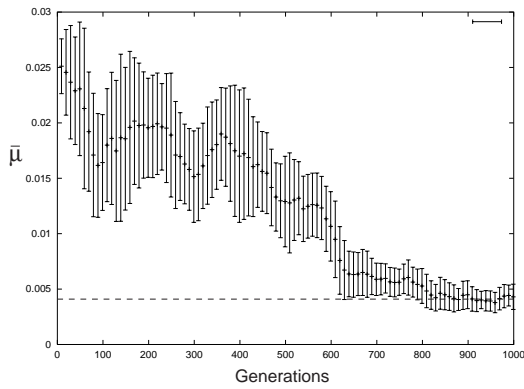


FIG. 1. Mean mutation rate evolving towards the optimal rate of $\mu_{opt} = 0.00445$. Error bars are one standard deviation about the mean. $\sigma = 5, \tau = 2, \nu = 25, N = 10^4$

Fig. 2 shows the evolution of mutation rates in detail in a population with a reasonably fast rate of change of mutation rates. This simulation has the same landscape parameters as Fig. 1. The mean mutation rate fluctuates around the optimum. For mutation rates close to the optimum fluctuations in selection are significantly larger than the selective advantages of one mutation rate over another. In this region the evolution of mutation rates is effectively neutral and thus the mean mutation rate conducts a random walk about the optimum. We also note that the population typically spends more time with mutation rates above the optimum than below. This is mainly a finite population size effect. When the peak moves and the population size is limited there is a relatively large probability that there will be no individuals representing a master-sequence with very low mutation rate on the new peak. This leads to a temporary increase in mutation rate in the population after an environmental shift.

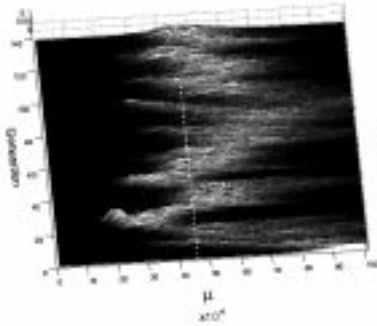


FIG. 2. Evolution of mutation rates of mutationally diverse population. $\mu_{opt} = 4.45 \times 10^{-3}$, $\sigma = 5, \tau = 2, \nu = 25, N = 10^4$

In nature the existence, and value, of an optimum mutation rate that results from a changing environment depends on many different parameters: the time between shifts in the selective environment, the complex structure of the fitness-landscape, the genome length, co-evolutionary effects, the strength of selection, neutrality in the fitness landscape and fluctuations due to finite population sizes etc. One must therefore be careful when comparing the results of a simple model, such as the one we have presented in this paper, and biological measurements. Nonetheless it is this range of possible differences between organisms and the complexity of their evolutionary environments that leads us to consider the possibility that simple laws of biology — such as the scaling of point mutation rates with genome length — are likely to have quite simple explanations that do not depend on the details of the particular organism. It is therefore worth comparing the results of the model presented in this paper with the biological data.

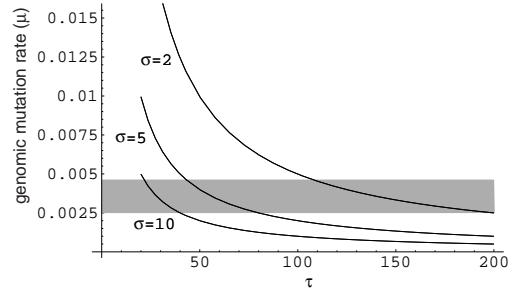


FIG. 3. The shaded region shows the genomic mutation rates for DNA based organisms listed in Table I. For low average fitness advantage σ the mutation rate is relatively insensitive to the frequency of changes in the environment. For clarity we have assumed $\langle k \rangle = 1$ in this figure.

For low mutation rates Eq. 9 is relatively insensitive to changes in the average fitness or size and frequency of environmental changes, as shown in Fig. 3. This insensitivity of the optimal genomic mutation rates to evolutionary parameters is important, since the bacteria and phages illustrated in table I are most unlikely to live in environments with the same types of time-dynamics and time-scales. In Fig. 3 we see that the sensitivity to one of the parameters in the model, σ or τ , depends strongly on in which region the other parameter is. For most realistic populations we may expect the selective advantage σ to be weak, maybe on average less than 2. The predicted mutation rate is then highly insensitive to the average time between shifts in the fitness landscape, e.g. $\sigma = 2$ gives $\tau \in [110, 200]$ for the organisms listed in Table I. It is also reasonable to assume the fitness landscapes of the organisms listed in Table I to be more similar to each other than to higher eukaryotes and since our predictions

as to Q_{opt} are rather insensitive to the details of $\sigma(t)$, τ and $h(\mu)$ we would expect many organisms to have approximately the same mutation rate per genome (within an order of magnitude). This is what we observe for simple DNA-based organisms.

VII. RNA VIRUSES

The lytic RNA viruses consistently show an extremely high mutation rate — orders of magnitude larger than that of any DNA viruses of similar size. This rate of around one substitution per genome per generation is inconsistent with the analysis conducted above for mutation rates evolving in a changing selective environment. Such high rates imply implausible values for the dynamic environment parameters.

As an explanation for the high mutation rates observed in many RNA viruses and the mutation rate scaling with genome length it has been suggested that these viruses have evolved the highest mutation rate possible to be able to adapt to a rapidly changing environment. The maximal mutation rate is then given by the error-threshold, which was first discussed in a model by Eigen *et al.* [18]. It basically states that on a singled peaked fitness landscape an organism must have high enough copying fidelity so that its relative superiority in reproduction rate multiplied by the probability of reproducing onto a perfect copy of itself must be larger than one, otherwise there will be no effective selection for the genotype. It has later been shown that the error-threshold can rather easily be generalized to include effects of a dynamic environment [19]. From this argument it is however not clear why RNA viruses should evolve towards the error-threshold while DNA based organism tend to have much lower mutation rates (by orders of magnitude). In this section we will combine the error-threshold with the model presented in this paper to suggest a possible explanation to the difference in observed mutation rates between DNA and RNA based organisms.

The dynamic environment model presented in this paper applies to organisms where the copying fidelity is encoded in a part of the genome that has little or no effect on fitness. In many viruses this may not be appropriate, partly because the proteins involved in mutagenesis may have a multitude of functions but also because the relatively high selective pressure towards short genome lengths will result in the overlap and multiple use of genetic material where possible. This give rise to a different possibility for the evolution of optimal mutation rates and might help explain the large differences between the observations for RNA and DNA based organisms.

We suggest that for organisms which have strong overlaps between genes coding for the mutation rate and genes coding more directly for reproductive advantage there is no effective selection for lower mutation rates, as long as the mutation rate is below the error threshold.

This argument is based on the assumption that most mutations are deleterious in terms of fitness, and that the relative fitness advantage on the local peak results in stronger selection pressure than the pressure towards lower mutation rates. We also assume that evolution of mutation rates usually affect regions of the genome where the organism need mutations to be able to adapt ot changes in the environment. If these assumptions apply we expect a population to have mutation rates close to the error-threshold. Changes to mutation rate is transient, assuming that the organism is not pushed beyond the error-threshold.

For this hypotheses to apply, viruses with high mutation rate (mainly RNA viruses) should have overlapping genes regulating mutation frequency as well as reproduction rate, whereas organisms with low mutation rates (such as those listed in Table I) should not have overlapping reading frames in their genomes. There are observations that support this, but it is unclear whether the correlation is strong enough for this hypothesis to be valid.

VIII. CONCLUSIONS

In this paper we have studied the evolution of mutation rates in a population of multi locus genomes. The genomic mutation rate μ_G leading to the greatest long term growth of a strain (the optimal rate) was analytically determined for reasonably general peak shifts and time-dependent replication rates $\sigma(t)$

$$\mu_G \approx \frac{\langle k \rangle}{\langle \sigma \rangle \langle \tau \rangle}$$

where $\langle k \rangle$ is the mean Hamming distance between successive fitness optima and $\langle \tau \rangle$ is the mean time between shifts. These optimal rates were quantitatively confirmed by computational simulations of populations whose mutation rates were allowed to evolve.

These continuous time multi-locus replicator models predict the kind of scaling of point-mutation rate with genome length that has been observed in some bacteria and viruses/phages and puzzled over for years. When combined with the consequences of the multiple use/pleiotropic encoding of copying machinery these models of the evolution of mutation rate in dynamic environments also suggest why lytic RNA viruses may have rates at or about the error-threshold.

We would like to thank Claes Andersson and Erik van Nimwegen for useful discussions. Thanks are also due to Mats Nordahl who has given valuable comments on the manuscript. Nigel Snoad and Martin Nilsson were supported by SFI core funding grants. N.S. would also like to acknowledge the support of Marc Feldman and the Center for Computational Genetics and Biological Modeling at Stanford University while preparing this manuscript.

-
- [1] J.W. Drake, B. Charelsworth, D. Charlesworth, and J.F. Crow. Rates of spontaneous mutation. *Genetics*, 148(4):1667–86, April 1998.
 - [2] J. Maynard Smith and E. Szathmáry. *The Major Transitions in Evolution*. W.H. Freeman, Oxford, 1995.
 - [3] M Kimura. On the evolutionary adjustment of spontaneous mutation rates. *Genet. Res.*, 9:23–24, 1967.
 - [4] E.G. Leigh. Natural selection and mutability. *Amer. Natur.*, 104:301–305, 1970.
 - [5] A.S. Kondrashov. Modifiers of mutation-selection balance: general-approach and the evolution of mutation-rates. *Genet. Res.*, 66:53–69, 1995.
 - [6] E.C. Cox and T.C. Gibson. Selection for high mutation rates in chemostats. *Genetics*, 77:169–184, 1974.
 - [7] E.C. Cox. Bacterial mutator genes and the control of spontaneous mutation. *Ann. Rev. Genet.*, 10:135–156, 1976.
 - [8] L. Chao, C. Vargas, B.B. Spear, and E.C. Cox. Transposable elements as mutator genes in evolution. *Nature*, 303:633–635, 1983.
 - [9] P.D. Sniegowski, P.J. Gerrish, and R.E. Lenski. Evolution of high mutation rates in experimental populations of *E. coli*. *Nature*, 387:703–705, 1997.
 - [10] G.T. McVean and L.D. Hurst. Evidence for a selectively favourable reduction in the mutation rate of the X chromosome. *Nature*, 386:388–392, 1997.
 - [11] E.R. Moxon, P.B. Rainey, M.A. Nowak, and R.E. Lenski. Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr. Biol.*, 4:24–33, 1994.
 - [12] E.R. Moxon and D.S. Thaler. The tinkerer’s evolving toolbox. *Nature*, 387:659–662, 1997.
 - [13] M. Radman. Enzymes of evolutionary change. *Nature*, 401:866–869, 1999.
 - [14] E.G. Leigh. The evolution of mutation rates. *Genetics Suppl.*, 73:1–18, 1973.
 - [15] J.H. Gillespie. Mutation modification in a random environment. *Evolution*, 35:468–476, 1981.
 - [16] K. Ishii, H. Matsuda, Y. Iwasa, and A. Sasaki. Evolutionarily stable mutation rate in a periodically changing environment. *Genetics*, 121:163–174, January 1989.
 - [17] J.H. Gillespie. *The Causes of Molecular Evolution*. Oxford University Press, Oxford, UK, 1991.
 - [18] M. Eigen and P. Schuster. The hypercycle. A principle of natural self-organization. Part A: emergence of the hypercycle. *Naturwissenschaften*, 64:541–565, 1977.
 - [19] M. Nilsson and N. Snoad. Error thresholds on dynamic fitness landscapes. *Phys. Rev. Lett.*, 84:191–194, 2000.